Monitoring national attainment standards

A collection of working papers

Paul E Newton

December 2008

Product code: Ofqual/08/3916

Contents

Preface

Towards the end of 2005, a small working group met within the Qualifications and Curriculum Authority (QCA) to consider methods of monitoring national attainment standards. To support our deliberations, I produced a series of working papers that consider what might be learned from examples of national monitoring systems around the world (including past experience in England) and explore issues to consider if a country were to set up a national monitoring system anew. Since these working papers were only intended to support internal deliberation, they were not published at the time.

On 14 October 2008, the Department for Children, Schools and Families (DCSF) announced major reforms to school accountability, including an end to compulsory national tests for 14-year-olds, noting that:

> A new expert group, made up of headteachers and education professionals, will advise on the details of the new arrangements. The group will also advise government on the introduction of national-level sampling at key stage 3 so that the performance of the education system as a whole can still be monitored by the public, year on year.[1]

Given this announcement, Ofqual decided to publish the working papers I prepared during 2005 and early 2006, while Ofqual was still the regulation and standards division of QCA. These are presented in the following pages, each one as an individual chapter.

No substantive changes have been made to these working papers since they were written. Consequently, where they refer to current developments internationally, they are likely to be out of date. However, most of the content is still relevant, and I hope that the collection may be of some use to the expert group in deciding their recommendations.

Paul E Newton
Head of Assessment Research
Office of the Qualifications and Examinations Regulator
December 2008

[1] http://www.dcsf.gov.uk/pns/DisplayPN.cgi?pn_id=2008_0229

Insights from England's Assessment of Performance Unit

The Assessment of Performance Unit (APU) was established in 1974 and ran until 1990. Although it seems to have acquired a reputation, amongst some, for having been England's 'best example' of a monitoring system, it didn't actually achieve this aim particularly well. Lessons to be learned from the APU experience, include:

1. There is no straightforward, uncontroversial solution to the problem of measuring change in the national attainment profile over time.

2. The development of a system for monitoring trends in the national attainment profile must not be rushed, else the baseline data (and, in all likelihood, data from following years) may prove useless. An effective national monitoring system cannot be brought 'on stream' in just a couple of years.

3. Purposeful, informed and consistent leadership is central to the success of a monitoring programme. Most importantly, its primary purpose must be made explicit from the outset, according to which detailed long-term plans must be made.

4. Modest designs – based on well-established principles and procedures – are to be preferred when constructing monitoring systems that are intended primarily to create robust trend results.

5. Even light sampling principles can translate into heavy sampling practice, if applied across numerous domains and if surveys are repeated frequently.

Introduction

Recent concern over the suitability of national curriculum test results for measuring change in the national attainment profile of pupils in England (for example Massey, et al. 2003; Tymms, 2004) has prompted discussion over the potential of alternative systems for fulfilling this function. In particular, the question 'why not bring back the Assessment of Performance Unit?' is frequently asked.

The Assessment of Performance Unit (APU) was established in 1974 within the Department for Education and Science (DES). It was originally headed by HMI Brian Kay, with terms of reference to:

> 'promote the development of methods of assessing and monitoring the achievement of children at school, and to seek to identify the incidence of underachievement.'

Although the terms of reference emphasised the identification of underachievement, it became apparent that the primary function of the APU would be to monitor trends in the national attainment profile more broadly.

By 1990, a total of 43 surveys had been administered, spanning five broad subject areas: language, mathematics, science, foreign languages and design and technology. The roll-out of the national curriculum, and its associated assessment system, signalled the eventual demise of the unit.

The purposes of this paper are to summarise the key features of the APU experience, to identify some of the problems that it encountered, and to draw out lessons that should be learned for the future.

Background to the formation of the APU

To understand the political significance of the APU, it is important to appreciate the context in which it emerged. Secondary education in England had recently undergone major structural reorganisation, with the move towards comprehensivisation and the associated decline of the 11+ examination. This was also a period – following publication of the Plowden Report – during which progressive educational ideas and child-centred education were becoming increasingly prominent. Concern over such significant reform was growing rapidly, particularly amongst traditionalists, who feared that standards in schools were falling. This culminated in the Great Debate on education during the early 1970s.

During 1972, the National Foundation for Educational Research (NFER) published results from the latest of a series of national reading surveys that had been in operation since the late 1940s and had been funded by successive education departments. The identification of a small fall in standards was seized upon, from one perspective, by traditionalists. However, from another perspective, the validity of the NFER results – and the survey methodology more generally – came under attack.

Shortly afterwards, the education secretary commissioned the Bullock Report into all aspects of the teaching of English and into the monitoring of standards in English. It was finally published in 1975, a year that also saw a high profile inquiry into the alleged failure of progressive education within William Tyndale school. The following year culminated in Callaghan's famous Ruskin College speech, with its emphasis on educational accountability.

Key facts about the APU

The formation of the APU had two principal precursors, both funded by the DES: first, the formation in 1970 of the Working Group on the Measurement of Educational Attainment; second, the commissioning in 1972 of the NFER to run feasibility projects under the title *Tests of Attainment in Mathematics in Schools*.

The establishment of the APU was announced, in August 1974, in the DES White Paper *Educational disadvantage and the needs of immigrants*. The unit would collaborate with the recently announced Educational Disadvantage Unit to develop criteria to improve the identification of educational disadvantage. In fact, this focus on

underachievement and educational disadvantage never featured prominently in the work of the APU: partly because of the lack of useful definition of underachievement, and partly because the survey approach adopted was not appropriate for focusing on relatively small sub-groups of the population. Others (such as Holt, 1981) suggested a more sinister explanation: the DES wished to gain some control of the curriculum, through backwash impacts of monitoring, but also wanted to conceal this intention by spinning the exercise as one of supporting the disadvantaged.

The unit oversaw 43 surveys of attainment, primarily of mathematics, language and science at ages 11 (end of primary school) and 15 (end of secondary school). Later surveys included foreign languages and design and technology, as well as the testing of 13-year-olds. The following table identifies when the 43 surveys were carried out.

Subject area	Age 11	Age 13	Age 15
Mathematics	1978-82, 1987		1978-82, 1987
Language	1979-83, 1988		1979-83, 1988
Science	1980-84	1980-84	1980-84
Foreign languages		1983-85	
Design and technology			1988

There were two main phases of APU operation. During the first phase, the test instruments, survey methodology and statistical procedures were developed; then a series of five annual surveys was administered. During the second phase, there was a sense of consolidation, during which further research was undertaken and further statistical analysis of the data collected so far. The second phase also involved the development of new instruments to be used in subsequent surveys, and the decision was made to move from an annual basis to a five-yearly basis.

The work of the unit was guided by three principal groups: the Co-ordinating Group, the Consultative Committee and the Statistics Advisory Group.

The **Co-ordinating Group** was a ‘technical and professional’ panel that included HMIs and members of DES, as well as representatives from the teaching profession, local education authorities, academia and the NFER. It had a particular focus on the assessment model and cross-curricular integration, as well as acting as a general sounding-board for APU ideas. It was charged with co-ordinating the working groups, monitoring teams and committees; however, in practice, it had less power than the other two principal advisory groups and was disbanded in 1980.

The **Consultative Committee** was a 'stakeholder' advisory group, including members of the teaching profession, local education authorities, academia, the business world and parents. The idea of this group was to ensure that all educational partners were kept 'on board' – something that the North Americans had failed to do when they introduced their National Assessment of Educational Progress (NAEP) in the late 1960s. The committee was charged with providing general advice to the programme. In practice, it acquired a power of veto (if not executive power).

The **Statistics Advisory Group** was a 'technical' group, made up of members from the field of statistics and assessment, from the NFER, DES and academia. It was charged with guiding the programme on matters related to statistics, sampling and item banking. It developed a particular focus on problems related to the assessment of trends over time and underachievement.

In addition to the three overarching groups, subject-specific **working groups** were established to take the work forward. These became **steering groups** once the monitoring work began in earnest. **Exploratory working groups** in personal and social development, aesthetic development and physical development were disbanded after it was decided these areas were not suitable for monitoring activity.

The development of assessment instruments was contracted out to subject-specific monitoring **teams**, as follows:

Mathematics	NFER
Language	NFER
Foreign languages	NFER
Science	Leeds University & Chelsea College (which later merged with King's College, London)
Design and technology	Goldsmiths College, London

Finally, the surveying process was contracted to the NFER Monitoring Services Unit.

The estimated programme running costs – per annum, from 1977 to 1981 – were as follows (see Gipps and Goldstein, 1983, p.187):

- £160,000 for central and administration costs
- £480,000 for the language, mathematics and science survey teams.

These have not been adjusted to present-day values.

Key events in the history of the APU

1970	**Working Group on the Measurement of Educational Attainment** established by the DES
1972	NFER commissioned to run feasibility projects: *Tests of Attainment in Mathematics in Schools* (TAMS)
1974 (August)	formal announcement of the newly established APU
1974 to 1975	internal consultation on development of model, by head of APU
1975 (June)	publication of *Monitoring pupils' progress* by head of APU
1975 (September)	first meeting of **Science Working Group**
1975 (October)	first meeting of **Language Working Group**
1975 (October)	establishment of the **Coordinating Group**
1976 (March)	visit by deputy director of NFER and head of APU to North America to study NAEP
1976 (May)	first meeting of the **Consultative Committee**
1976 (October)	first meeting of **Mathematics Working Group**
1976 (October)	first meeting of **Personal and Social Exploratory Group**
1977 (January)	first meeting of the **Statistics Advisory Group**
1977 (June)	first meeting of **Aesthetic Development Exploratory Group**
1977 (June)	first meeting of **Physical Development Exploratory Group**
1977 (December)	first meeting of **Foreign Language Working Group**
1978 (March)	first **primary mathematics** survey (reported in Jan 1980)
1978 (October)	first **secondary mathematics** survey (reported in Sept 1980)

1979 (March)	first **primary language** survey (reported in Sept 1981)
1979 (Oct/Nov)	first **secondary language** survey (reported in Mar 1981)
1980 (March)	first **primary science** survey (reported in Dec 1981)
1980 (November)	first **secondary science** survey (reported in Dec 1982)
1989 (March)	decision not to undertake any more surveys
1989 (April)	supervision of remaining teams taken over by Evaluation and Monitoring Unit of SEAC
1989 (September)	introduction of the new national curriculum in schools

Key features of the APU assessment model

The approach to assessment adopted by the APU was explicitly modelled on the USA's National Assessment of Educational Progress, which had been developed during the 1960s and was first surveyed the year prior to the formation of the Working Group on the Measurement of Educational Attainment. The key features of this model included:

- **light sampling** – selecting only a small number of schools for each survey, and assessing only a few pupils from those school
- **matrix sampling** – testing different pupils on different sub-tests/tasks, so as to assess a much wider range of knowledge, skill and understanding than would be possible using a single test, and to reduce the assessment burden on individual pupils; typically, each survey would involve around 20-30 different sub-tests/tasks
- **anonymity of results for schools and pupils** – results would only be reported at the national level or to compare sub-groups defined by gender, ethnicity and such like, ensuring that the assessment remained low stakes for pupils and schools thereby reducing the likelihood of teaching to the test, so as not to risk negative curriculum washback
- **non-release of questions** – the specific questions used in the survey would remain confidential – again, reducing the likelihood of teaching to the test, so that the apparent difficulty of the questions should not change over time.

All but two of the surveys included schools from England, Northern Ireland and Wales. For each survey, around 10,000 pupils were selected from England (about one per cent) and around 2,500 pupils were selected from Wales and Northern Ireland, respectively (about six per cent) The sample sizes were not calculated with great scientific precision and the science team was particularly concerned at this aspect of the process.

All teams were committed to developing instruments that emphasised:

- **communication**, in various formats
- interpreting and solving **purposeful problems**
- **practical situations** and realistic contexts
- **main ideas** and basic concepts of each subject.

Each subject assessed the following learning outcomes:

- concepts and skills
- problem solving strategies
- attitudes.

These foci were different from the traditional testing models of the time (which were far more content- than process-driven) and the emphasis given to process skills explicitly reflected the monitoring purpose. That is, it was explicitly intended to focus on aspects of each subject domain that were least likely to change in relevance over time (in the way that specific content typically does). The emphasis on process skills also reflected the fact that there was no common curriculum that all pupils were following. In reality, it is impossible to design an assessment that is devoid of content and context and reflects only process skills, but the teams were committed to keeping process skills to the fore as much as possible.

By way of illustration, the assessment formats for the original mathematics surveys took the following form:

- **written tests** administered by sample pupils' teachers, with pupils responding in writing to written questions.
- **practical tests** administered by experienced teachers who had been recruited and trained by the team (around 30 per survey), with pupils responding orally to orally presented questions

- attitudes evaluated by **questionnaire** and by **assessor judgement** of practical performance.

The maximum testing times for primary and secondary pupils were originally recommended to be 45 minutes and one hour, respectively. However, all teams found that it was useful and acceptable to exceed these guidelines – sometimes substantially, particularly for practical assessments.

Marking was generally undertaken by experienced teachers who had been recruited and trained by the teams, although some teams employed students to mark questions that could be objectively marked. All marking had some quality control checking while writing scripts and oracy tapes were impression-marked twice.

Unlike present national curriculum tests, the analytical focus was on performance over time at the sub-domain level of each subject, rather than at the full-domain level. Thus the intention was to explore change at the level of, for example number, measures, algebra, geometry, probability and statistics rather than at the level of mathematics overall. These sub-domains were precursors to national curriculum attainment targets.

The analysis of results was particularly complicated for two principal reasons. The first challenge was the complexity simply attributable to the matrix sampling model, whereby different pupils attempted different sub-tests/tasks. This meant that techniques had to be developed to combine the results in an appropriate fashion.

The principal analytical challenge, which the unit never satisfactorily solved, was the measurement of change in attainment over time. The original intention was to use **item response theory** (IRT) techniques. These would, in theory, enable all pupils and questions to be calibrated to the same scale, even though different pupils would take different sub-tests/tasks. It would also, in theory, allow the teams to discard and replace questions that appeared to become out-of-date over time. In practice, these techniques came in for serious criticism; so much so that the original aim of monitoring change in attainment over time was effectively absent from the second phase of surveying. The science team adopted a slightly different analytical technique: **generalizability theory**. However, it came under attack for essentially the same reasons as IRT did.

The major success of the APU was not in monitoring attainment standards, but in the development of innovative assessment materials (which strongly influenced the nature of national curriculum assessment) and in the detailed study of performance characteristics (with important lessons for teachers, academics and test developers alike). Key reporting foci included:

- pupils' errors, misconceptions and alternative conceptual frameworks

- the impacts of task context, task purpose and task presentation on performance
- comparative performance between sub-groups based on age, gender, attainment bands – these relative differences between sub-groups within years became the key focus for policy makers as the attempt to estimate absolute change for the population was abandoned.

APU publications included:

- **reports** on individual surveys
- **review reports** on the first phase surveys
- **topic booklets** for teachers
- **topic leaflets** for teachers
- practical **assessment kits**
- (nine) **newsletters** for schools and LEAs.

As a final point it is worth noting that, while the APU had many noble plans, it was not always possible to realise them. For example, it was originally intended to assess a very broad range of educational attainments, including ethical, aesthetic and physical domains. However, these were eventually considered to present too significant an assessment obstacle to pursue. In addition, the original intention was that assessment should be based on a cross-curricular model, rather than being confined within the traditional subject domains. This aspiration was essentially unfulfilled, and was not helped by the teams working largely independently. Not all of the grand designs for the assessment procedures were realised either, despite the considerable advances that were achieved. For instance, there was an early intention to rely heavily upon 'open-ended' writing questions to stretch the ablest; but the apparent difficulty of developing such questions and the costs of marking them challenged this intention.

Major criticisms of the APU

The following criticisms of the Assessment of Performance Unit have been taken primarily from Gipps and Goldstein (1983).

Rationale

Apart from the general confusion between the aims of identifying underachievement and of monitoring national attainment trends, there was a deeper confusion as to whether the APU exercise was primarily about monitoring or about research. The early surveys generated relatively bland findings – which the unit insisted should be written up without elaboration or interpretation. For instance, more 11-year-olds in the

south can do simple fractions than in the north. These findings were not of a great deal of relevance, either to the public, to teachers or to policy makers. The in-depth studies, mooted in the early days (as necessary for exploring possible causal links) failed to materialise. Subsequent reporting was allowed to be more interpretative and, hence, more interesting and potentially useful.

Management

The management of the APU was felt to be weak, with a general lack of long-term planning, no central co-ordination of the work of the teams, and confusion between the roles of certain advisory groups.

There seemed to be a lack of recognition that the programme would require technical expertise in research methodology, as distinct from expertise in statistical methodology, showing a lack of experience in running large-scale research projects.

There were frequent changes of personnel within the unit, resulting in a lack of ‘collective memory’. And it was never made clear to ‘outsiders’ that the APU was a long-term project that would take many years to come ‘on stream’.

Practical and technical problems

Numerous practical issues caused problems for the unit. One important problem was that the light sampling of schools – which were made in preparation for the first mathematics survey – turned into moderately heavy sampling by time the other subjects were being assessed as well. The NFER noted that, by 1979, around a third of all maintained secondary schools had taken part in some APU work (despite only having assessed maths twice and language once). This resulted in some pressure to reduce the frequency of testing.

The decision not to measure home background variables eventually reduced the power of the monitoring instrument for creating theories of attainment. Again, this revealed confusion over whether the primary purpose of the unit was one of research or monitoring.

As time went on there was an increasing problem of non-response at school and individual pupil level. When combined with evidence from NFER analyses that ‘chased’ schools/pupils tended to perform worse on average, this suggested that non-response was likely to bias the results, more so each year as the non-response increased.

The most significant technical criticism concerned the two main assumptions of the Rasch model, which formed the basis of the item response theory techniques – on which the majority of the statistical analysis was based. The basic assumptions of the Rasch model are that:

- each pupil can be said to have a fixed level of attainment in a subject domain (or, more plausibly, sub-domain), which operates in the same way for all items that the pupil attempts
- each item can be said to have a fixed level of difficulty, which operates in the same way for all pupils who attempt it in each test that it appears, each year it is administered.

Obviously, these assumptions are false when interpreted literally; the issue is whether they are implausible even when interpreted probabilistically. Harvey Goldstein, a leading professor of educational statistics, argued strongly that they were implausible in an educational context and, moreover, that proposed technical 'solutions' to assumption violations were untenable.[2] The assumptions are particularly challenged when different pupils study different syllabuses in different schools, and when syllabuses change over time – precisely the contexts in which the APU was operating. Although tests may be developed whose items appear to meet the strict assumptions, these are unlikely to be sampling the domain fairly and, as such, are likely to be educationally invalid.

Although the NFER continued to use Rasch analyses for linking standards between discrete sub-tests/tasks within years, it eventually abandoned them for comparisons between years, accepting that the constant relative difficulty assumption was invalid. The APU never agreed upon a satisfactory solution to the problem of measuring trends in national attainment over time.

Dissemination

During the early years of the APU, the distribution of reports was poor. Few teachers or education authority advisers read the full reports, although many teachers did read the NFER summaries. As the programme progressed, dissemination improved, partly due to the production of more useful content and partly due to more accessible formats.

Timing

The mathematics group, in particular, felt that they were continually being rushed: their requests for a delay in the monitoring programme were rejected; their desire for three pilot surveys was realised as only one; they experienced a high turnover of staff

---

[2] For example: '… there is an inherent flaw in the item bank concept which would make it unworkable in practice. If we suppose that each of the items in the bank has a prescribed difficulty value, then it is strictly meaningless within the context of the Rasch model to speak of one item being more applicable to one point in time rather than another… an item bank which is designed so that out-of-date items can be replaced is a strictly non-Raschian concept." (Goldstein, 1979, p.217)

and a resulting shortage of personnel. The constant rush meant that there was no time for identifying and remedying problems experienced in the first year of testing.

In fact, all three teams suffered from a rapid turnover of staff, put down to the constant pressure of work combined with a lack of opportunity to ‘side-track’ into interesting research issues.

Gipps and Goldstein (1983, p.163) summarised the effectiveness of the first phase of the operation of the APU as follows:

> 'To sum up then, the conclusion on the APU’s progress must be that it has had partial success. It has succeeded in test development, it has persuaded LEAs to co-operate in the surveys, and it has made a start on the study of circumstances in which children learn, although the outcome here is uncertain. On the other hand it has failed on underachievement and has not yet had any success in describing changes in performance over time.'

Lessons from the APU experience

The following ‘lessons to learn’ represent a personal take, stemming from a reading of the relevant literature in the context of present-day concerns.

The most important lesson to learn from the APU experience is that there is no straightforward, uncontroversial solution to the problem of measuring change in the national attainment profile over time. The APU explored two sophisticated techniques, based on item response theory (IRT) and generalizability theory, respectively, and found neither to be satisfactory.

Interestingly, as the APU dropped the use of IRT for ensuring comparability of standards over time, so contractors in the United States were exploring its potential. Today, in the States, the use of IRT by test development agencies is routine and it is routinely employed in analyses for NAEP. Although the sophistication of IRT modelling is far superior nowadays, it is still essentially based on the same assumptions as those of the original Rasch model. These do not always hold, as problems in applying IRT to NAEP have revealed (e.g. Beaton and Zwick, 1990). Whether such modelling techniques are more suitable in the present UK context than they were two decades ago is worthy of consideration. However, it would be wrong to assume that the necessary statistical solutions can be found easily. The required technical debate would be needed before any decision is made concerning the establishment of a new unit akin to the APU.

This introduces another important lesson: timing. The time-scale according to which the APU was introduced was not short. The first surveys were administered eight years after the establishment of the original Working Group on the Measurement of Educational Attainment, and four years after the establishment of the APU. Reports on the first mathematics surveys were not available until over three years after the

initial meeting of the mathematics working group, even though many of the instruments had already been developed during earlier feasibility studies. Time pressure was felt acutely. This resulted, for mathematics, in problems such as insufficient piloting and not being able to remedy identified problems. During any attempt to monitor standards over time, the most important concern is to get the methodology and instruments right from the outset – otherwise the baseline data will be useless – so the temptation to rush simply has to be avoided at all costs.

The APU experience made clear that purposeful, informed and consistent leadership is central to the success of a monitoring programme. It needs to be clear from the outset exactly what is intended to be achieved, and those managing the programme require insight into what would be needed in order to bring about the intended end. The programme managers – and not simply the contractors and advisers – need some grounding in the principles of research project management, ideally as applied to educational assessment. Furthermore, since consistency over time is the fundamental principle underlying the measurement of change, this should be reflected in the management structure and (where at all possible) personnel. Also central to the success of a monitoring programme is detailed long-term planning from the outset.

The APU illustrated that grand designs cannot always be realised. This was not only true in relation to the application of complex new statistical techniques; it also proved true in relation to the assessment of some of the more abstract attainment constructs. Modest designs – based on well-established principles and procedures – are to be preferred when constructing monitoring systems. The development of a monitoring system is not the appropriate context for piloting novel assessment techniques. Again, the basic requirement of a monitoring system is that you get it right in the first year, else the baseline data will be useless. Novel assessment techniques need many years of piloting before their first administration; if sufficient time for extensive piloting is not available, then tried and tested approaches should be used.

A final important lesson from the APU is that even light sampling principles can translate into heavy sampling practice, if applied across numerous domains and if surveys are repeated frequently.

Principal bibliography

Three texts were particularly important in compiling this report:

Holt, M. (1981). *Evaluating the evaluators*. Hodder and Stoughton. London.

Gipps, C. and Goldstein, H. (1983). *Monitoring children: an evaluation of the Assessment of Performance Unit*. Heinemann Educational Books. London.

Foxman, D., Hutchison, D. and Bloomfield, B. (1991). *The APU experience: 1977 – 1990.* School Examinations and Assessment Council. London.

Additional references

Beaton, A.E. and Zwick, R. (1990). *The effect of changes in the national assessment: disentangling the NAEP 1985-86 reading anomaly*. Educational Testing Service. Princeton, NJ.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5(2), 211-220.

Massey, A., Green, S., Dexter, T. and Hamnett, L. (2003). *Comparability of national tests over time: key stage test standards between 1996 and 2001*. QCA. London.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, 30 (4), 477-494.

Insights from the USA's National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) is probably the world's most sophisticated and well-researched instrument for monitoring attainment trends. It assesses both long-term trends and short-term trends although, crucially, it uses separate instruments to do so. NAEP offers many insights into the design and operation of a national monitoring system, of which the following are a selection.

1. The effective design of a national monitoring system is a complex project, which is likely to take many years. Planning for NAEP began in 1963, but the first surveys were not administered until 1969. The development period was considerably longer than policy makers had originally anticipated.

2. NAEP officials initially aspired to operate a highly valid and inclusive model of assessment. In practice, it turned out to be less inclusive (since it excluded certain groups) and less valid (since it emphasised traditional pencil-and-paper formats) than anticipated.

3. NAEP officials originally intended to report results on a task-by-task basis, to provide messages of particular relevance for the curriculum and for pedagogy. Results came to be reported primarily at the overall subject level, which oriented them more toward education policy than teaching practice.

4. Despite extensive ongoing funding for research, development and evaluation, NAEP has not been immune to technical problems associated with the assessment of attainment and the measurement of change. Throughout its history, major technical anomalies have been encountered, for example the 1986 reading anomaly, and major technical controversies have arisen, such as the 1993/1999 standard setting controversy.

5. The US Department of Education is presently sponsoring a considerable amount of research into the potential of e-assessment to support and enhance NAEP's assessment model. However, there is no intention to implement any innovation that has not been thoroughly evaluated (with respect to validity, reliability and comparability) or costed.

Introduction

The National Assessment of Educational Progress (NAEP) is known in the USA as the Nation's Report Card. It has two central objectives:

1. to measure student progress over time

2. as educational priorities change, to develop new assessment instruments that reflect current educational content and assessment methodology.

It has monitored trends in attainment standards since the late 1960s, although its history has not been entirely trouble-free.

During the early years of NAEP, one of its major challenges was to overcome the widespread perception that it was intended as a Trojan Horse to achieve central control of the curriculum. This fear was to some extent dispelled by a decision to grant responsibility for survey development and administration to the Education Commission of the States (ECS). However, many have remained concerned about possible negative consequences for curriculum and pedagogy in the US, particularly given the continuous advance of the accountability movement and the spectre of increasingly high stakes for NAEP results.

The most serious technical challenge faced by NAEP came to light a couple of years after responsibility for survey development and administration was transferred to contractors: the Educational Testing Service (ETS). They had won the contract partly on the basis of proposals to reconfigure NAEP so that it would be able to provide more useful results. This required the introduction of complex statistical modelling techniques, based on a principle known as Item Response Theory (IRT). Unfortunately, the transition to the new methodology resulted in implausible results for some of the early surveys; particularly the transition for reading from 1984 to 1986, which came to be known as the NAEP reading anomaly. The anomaly was largely attributed to a failure of the assumptions required for IRT to hold true in practice (see Beaton and Zwick, 1990). One crucial lesson was hammered home by the reading anomaly: 'When measuring change, do not change the measure.' (Beaton, 1990, p.165).

Despite this, if the decision *is* made not to change the measure over time, then it will progressively become less and less relevant to the present day. Since NAEP had used largely the same assessment frameworks, instruments and procedures from the late 1960s to the mid-1980s, this problem of decreasing relevance was beginning to become obvious. Increasing tension between the dual aims of measuring change and maintaining relevance led to a decision to split the national assessment into two discrete components:

1. Long-term trend NAEP

2. Main NAEP

Long-term trend (LTT) NAEP – the original model – would continue to assess educational progress in reading, writing, mathematics and science using largely the same assessment frameworks, instruments and procedures as had been used since the beginning. This would enable the measurement of trends over the long-term. In

contrast, Main NAEP would be designed to evolve frameworks, instruments and procedures, to monitor attainment based upon 'state of the art' conceptions of curriculum, pedagogy and assessment. Nowadays, Main NAEP frameworks, instruments and procedures are kept constant for a period of around 10 years, across which it is possible to monitor short-term trends. Results from the Main and LTT trend lines are officially recognised not to be comparable.

With an increasing focus on accountability, one of the most important of recent innovations has been the development of two spin-offs from Main NAEP:

1. State NAEP

2. (Trial) District NAEP

Both State and District NAEP are based on the same assessments as used for Main NAEP. The difference is simply that additional samples are drawn for the participating states/districts to allow reliable trend inferences to be drawn at these levels. Today, the most frequent surveys are the biennial Main/State NAEP surveys in reading and mathematics, tested in grades 4 and 8. Surveys in other subjects occur less frequently.

Key events in the history of NAEP

1963 US Commissioner of Education, **Francis Keppel**, instigates a project to explore options for reporting on the condition and progress of American education, including two preliminary conferences.

1964 **Exploratory Committee for the Assessment of Progress in Education** (ECAPE) is established by the Carnegie Foundation, with responsibility for designing an appropriate assessment system.

1965 A **Technical Advisory Committee** (TAC) is appointed to support the work of ECAPE.

1968 US Office of Education begins to contribute funding, with responsibility for oversight of NAEP resting with the **National Center for Educational Research** (NCER).

1968 ECAPE is promoted to **Committee for the Assessment of Progress in Education** (CAPE).

1969 TAC is promoted to **Analysis Advisory Committee** (ANAC) and an **Operations Advisory Committee** (OPAC) formed.

1969 Responsibility for survey development and administration is transferred to the **Education Commission of the States** (ECS), funded initially by grant and later by contract.

1969 ECS creates the **National Assessment Policy Committee.**

1969 First **LTT NAEP** tests in **science** are administered.

1971 US Office of Education transfers responsibility for oversight of NAEP, from the National Center for Educational Research to the **National Center for Education Statistics** (NCES).

1971 First **LTT NAEP** tests in **reading** are administered.

1972 US federal government solely funds the project (to the tune of $4.5 million) for the first time.

1973 First **LTT NAEP** tests in **mathematics** are administered.

1978 Congress enacts legislation to transfer sponsorship of the programme from NCES to the **National Institute of Education** (NIE) and to create an **Assessment Policy Committee**, with responsibility for design and validation, to be appointed by the contractor (ECS).

1982 Willard Wirtz and Archie Lapointe publish a major critique of the design, administration and impacts of NAEP.

1983 ***A Nation At Risk*** – major critique of US education – is published.

1983 Responsibility for survey development and administration is transferred from ECS to the **Educational Testing Service** (ETS), resulting in a major redesign of NAEP.

1983 A new **Technical Advisory Committee** is appointed.

1984 First **Main NAEP** tests are administered under new model.

1987 Report of the **Alexander-James Commission** on the future of NAEP is published.

1988 Congress agrees legislation to extend the number of subjects assessed; to authorise **Trial State NAEP**; to form the **National Assessment Governing Board** with responsibility for steering and supervising the conduct of NAEP; and to increase NAEP funding (authorising annual budgets of at least $11.5 million in 1989, to almost $20 million in 1993).

1988 Technical Advisory Committee is renamed as the **Design and Analysis Committee.**

1990 First **Trial State NAEP** tests are administered (along the lines of Main NAEP, with additional samples drawn for participating states).

1990 Achievement levels for reporting Main NAEP results are established: basic, proficient and advanced.

1996 Word 'trial' is dropped for State NAEP.

2001 ***No Child Left Behind* Act** requires biennial State NAEP assessment for all funded states, at grades 4 and 8 in reading and mathematics (to commence from 2003).

2002 First **Trial Urban District NAEP** tests are administered (along the lines of Main NAEP, with additional samples drawn for participating districts).

2004 Significant changes are made to assessment instruments for LTT surveys for reading and mathematics (with 'bridging' studies to link the trend lines).

NAEP management and organisation

NAEP is a very large programme, supported by multiple agencies and contractors working in close collaboration. The programme is owned by the NCES, within the Institute of Education Sciences of the US Department of Education.

In 1988, the National Assessment Governing Board (NAGB) was appointed by the secretary of education as a largely independent body responsible for setting NAEP policy and for developing the framework and test specifications that serve as the blueprint for the assessments. The members of NAGB include governors, state legislators, local and state school officials, educators, business representatives and members of the general public.

The 2003-2006 assessments are contracted out to many organisations, including the following:

1. alliance coordination (Educational Testing Service (ETS))
2. design, analysis, and reporting (ETS)
3. question development (American Institutes for Research (AIR))
4. materials preparation, distribution, and scoring (Pearson)
5. sampling and data collection (Westat, Inc.)

6. web operations and maintenance (Government Micro Resources, Inc. (GMRI))

7. dissemination and outreach (Hager Sharp)

8. state service center (Westat, Inc.)

9. NAEP state coordinators (individual state education agencies)

10. state analysis (AIR)

11. meeting logistics (Hager Sharp)

12. NAEP quality assurance (Human Resources Research Organization, HumRRO).

(See http://nces.ed.gov/nationsreportcard/contracts/procurements.asp for further details on NAEP partners and their responsibilities.)

The NAEP programme also appoints many other contractors to undertake research and evaluation activities.

Key features of the NAEP assessment model

In common with many other national educational monitoring systems, NAEP operates on a number of key principles, including:

- a prime focus on the core subject domains of reading, writing, mathematics and science, with surveys in additional subject domains from time-to-time
- a survey schedule no more frequent than biennial
- the sampling of a relatively small number of students
- the sampling of a relatively large number of questions/areas from the subject domains assessed
- the administration of assessments under controlled conditions, by trained staff
- performances evaluated by trained markers
- no reporting of results for students or schools, so the assessment is low stakes for these participants.

The first major NAEP reorganisation began in 1983, when ETS won the contract for survey development and innovation. Whereas LTT NAEP was to remain targeted at 9-, 13- and 17-year-olds, Main NAEP began assessing pupils in grades 4, 8 and 12. Also, while results from LTT NAEP were initially analysed and reported question-by-question, ETS ensured that both Main and LTT NAEP would report in terms of scale

scores, which presented overall average results for each domain, to enable trends to be reported at the subject level.

During the late 1990s and early 2000s, it became clear that maintenance of the LTT surveys was becoming increasingly problematic. In 1999 the writing survey was deemed too unreliable to be continued; by 2005, it had been decided that the science survey – which was too out-of-date to be continued – should also be brought to an end. From 2004 onwards, the LTT surveys have only operated for mathematics and reading. Unfortunately, dropping writing and science meant that administration procedures had to be changed, since the booklets all contained a mix of questions from across the subject domains. This gave an opportunity for making a number of additional changes, to bring features of the assessment process up to date while retaining the underlying assessment frameworks. Special 'bridging' studies were required to model the impact that these changes might have on the validity of the LTT results.

Further details on the NAEP assessment model are presented in Annex 1.

Insights from the NAEP experience

With a programme as large and enduring as NAEP, no doubt many important insights could be drawn. Three different kinds of insight are presented below, with a particular focus on lessons for policy makers in England, should they consider implementing a similar monitoring system in the future.

Retreat from early aspirations

The process of development ended up taking much longer than had been intended, due to the many unexpected difficulties that were faced in developing such large suites of novel instruments (Vinovskis, 1998).

Numerous other early aspirations were retreated from as their implications became apparent and as priorities changed. Many of these were highlighted in an article by Jones (1996), as summarised below.

- Whereas the original intention was to assess knowledge that could be obtained from any source, assessment objectives ended up being limited to those that were generally taught in schools.
- Despite an intention to sample every group that could possibly be assessed, a number of groups came to be excluded, including students with severe sensory handicaps and students not in school.
- Despite an intention to develop a full range of assessment formats (particularly complex, authentic, performance assessments) fewer innovative question formats were developed and, in fact, with the appointment of ETS as contractor

in the mid-1980s, an even greater proportion of multiple-choice questions were produced.

- Whereas the original intention was only to publish results on a task-by-task basis, results began to be reported by groups of exercises and, ultimately, with the appointment of ETS as contractor, results started to be published at the overall subject level.

- Finally, with the development of Main (as distinct from LTT) NAEP, the survey began to be seen as an agent of change, and exercises began to be focus on desired rather than present curricula – this contradicted one of the core aspirations, which was not to have a 'washback' impact.

One of the most significant changes was the move from task reporting to subject reporting. Nowadays, results for all versions of NAEP are presented predominantly at the overall subject level (at the national/state/district level and also broken down by various sub-groups). This might be seen as a change of emphasis from the use of results for identifying messages for curriculum and pedagogy (professional messages) to the use of results for identifying differential performance (political messages). In large, this change of emphasis was due to the fact that too few professionals had (or had made) the time to explore the detailed information contained in the volume of task analyses of the early years.

Technical problems encountered

Technically speaking, NAEP is a very complex enterprise. It has had to face, and attempt to overcome, some particularly challenging problems over the years. The following paragraphs highlight two of the most high-profile of challenges to the validity of NAEP results: one related to the long-term trend, and one related to standards in the main survey.

Measuring change in the national attainment profile over time presents many challenges. When results are reported on a task-by-task basis these challenges are kept to a minimum. As long as the same tasks are repeated in the same way, using comparable samples of students, then useful inferences can be drawn. However, there are still two major problems. First, it is unclear how far to generalise inferences from changing performance on specific tasks, to changes in broader features of attainment or proficiency. In short, it is hard to tell whether the observed changes on tasks are educationally significant. Second, it is unclear the extent to which changing performance on specific tasks should be attributed to the students (decreasing attainment or proficiency, for example) or to the tasks (for example decreasing relevance).

A principal benefit of IRT modelling is that it should help to ameliorate both of these problems, by calculating a level of difficulty for each task on the same scale as all

other tasks. This enables subject-level, rather than task-level, reporting (general rather than specific inferences); and it allows outdated tasks to be replaced by new ones (that assess the same aspect of attainment at the same level of difficulty). Unfortunately, IRT requires some very strong assumptions about the nature of task difficulty and student attainment, both of which are often violated in practice. The more the assumptions are violated, the more meaningless the reported results will be. However, even when assumptions have been seriously violated – potentially rendering the results meaningless – results will still tend to emerge from the IRT technology. Furthermore, there may be no clear evidence from which to conclude that error has occurred, so the error may remain undetected.

The NAEP reading anomaly was only inferred in response to evidence of apparent changes in attainment that seemed simply implausible. The investigation into what might have gone wrong was extensive and didn't report for a few years (Beaton and Zwick, 1990). Even then, it was not entirely clear what had gone wrong, but the general conclusion was that the same test questions had functioned differently (i.e. had become more or less difficult) when they had been presented in different orders and contexts, and with the impact of a variety of other apparently trivial changes to administration. It was concluded that: 'Changes in trend assessment methodology are fraught with danger and should be undertaken only with great care.' (Beaton, 1990, p.165). This is a useful warning; note that each time a task is retired from the pool, and a new one introduced, this represents a change whose impact might not be trivial.

In addition to challenges faced in monitoring trends over time, NAEP has faced a number of other technical problems. One of the most high-profile, and controversial, of debates of recent years has concerned the establishment of achievement levels on Main NAEP. Achievement levels were introduced in 1990 to identify the proportions of students at each grade who could be classified as either below basic, basic, proficient or advanced. To make these classifications, the score scale must be divided into achievement bands. In a sense, these bandings are arbitrary; and they require the exercise of subjective judgement. The key question is whether the resulting achievement levels are either too arbitrary, or too subjective, to be defensible.

A major evaluation published in 1993 concluded that NAEP achievement levels were indeed too arbitrary and subjective to be defensible; in particular, the report concluded: 'the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make' (Shepard, et al. 1993, p.77). A subsequent evaluation published in 1999 reached the same conclusion: 'NAEP's current achievement-level-setting procedures remain fundamentally flawed. The judgment tasks are difficult and confusing; raters' judgments of different item types are internally inconsistent; appropriate validity evidence for the cutscores is lacking; and the process has produced unreasonable results.' (Pellegrino, et al. 1999,

p.182). Conclusions such as these have certainly been challenged (for example by Hambleton, et al. 2000) yet the controversy over standard setting remains unresolved. Recently, two key figures within the NAEP community have concluded: 'The standards-setting movement is marching ahead. At this point, the policy demand to set standards may be ahead of the technology resources to set them.' (Loomis and Bourque, 2001, p.214).

The future of NAEP

Legislation exists to ensure that NAEP will continue to operate for the foreseeable future. Following the 2004 revisions, LTT NAEP should remain relatively stable; while Main NAEP will continue to evolve as its remit requires. This will involve inevitable changes of framework, but will also include changes in response to developments in the technology of assessment. As just discussed, it is likely that procedures for standard setting will evolve in response to new research and theory. Perhaps the most significant aspects of change in the foreseeable future are likely to concern the incorporation of technology related to e-assessment.

Present NAEP technology already makes some use of e-assessment – the on-screen marking technology developed by Pearson, for example. However, as indicated in Duran (2003), there are further possibilities, including:

1. computer-based presentation of items, and recording of responses, on assessments of existing NAEP constructs
    - assessment of skills not open to paper-and-pencil testing
    - computer adaptive testing
2. extension of NAEP to assessments of new constructs (for example word processing as a sub-domain of writing, internet searching as a sub-domain of reading)
3. computer enhancement of assessment processes
    - computerised item development
    - computerised scoring of responses
    - computerised distribution of results in new media

Although there is clearly interest in the possibilities afforded by e-assessment, there is no desire to rush progress. As noted in the introduction to a recent evaluation document: 'embracing new technologies does not mean NAEP should rush to use every new technology in operational assessments. On the contrary, in its position of leadership, NAEP must thoroughly evaluate new technologies to address both

validity and cost issues and introduce them to operational NAEP only when these issues have been addressed' (Duran, 2003, page 4).

Annex 1 Key features of LTT and Main NAEP

	LTT NAEP	Main NAEP (plus information on State NAEP)
Subject domains assessed	Reading and mathematics – writing and science have recently been discontinued for technical reasons.	Science, reading, mathematics and writing are the principal subject domains. U.S. history, civics, geography and the arts have also been assessed during the past decade or so, while assessments for economics, world history and foreign language are under development. (State NAEP is administered only for science, reading, mathematics and writing.)
Assessment frameworks	Remained largely unchanged since the first administrations. Original frameworks tended to be structured around subjects as studied at school. They were not linked to specific curricula, but objectives did end up being tied to a broad status quo of school curricula.	Periodically revised to reflect the 'state of the art' in terms of curriculum, pedagogy and assessment. Frameworks and instruments remain constant for around a decade, to monitor short-term trends. The frameworks are revised using curriculum experts, policymakers and members of the general public. They decide what aspects of, and how, a particular subject ought to be assessed.

	LTT NAEP	Main NAEP (plus information on State NAEP)
Student groups	Students aged 9, 13 and 17 from public and private schools. Low participation rates have occasionally prevented the analysis of results from private schools.	Students from grades 4, 8 and 12 from public and private schools. (State NAEP assesses only grades 4 and 8 and samples pupils from public schools only.)
Survey frequency	Administered every four years.	Main (and State) NAEP administered every two years for reading and mathematics (for all states with Title I funding). Main (and State) NAEP administered every four years for science and writing for all states. Other subjects are assessed periodically.
Question sampling and presentation model	LTT NAEP was originally based on simple multiple matrix sampling, in which: ■ the questions were divided into discrete booklets ■ all students in a particular session responded to the same booklet, paced by an audiotape lasting around 45 minutes ■ students responded to questions on a range of subject areas. Post-2004, the following	Main NAEP is based on balanced incomplete block spiral (BIB-spiral) multiple matrix sampling, in which: ■ questions are divided into blocks of similar length ■ blocks are assembled into booklets, such that each block appears in the same number of booklets, every pair of blocks of a certain type appears together in at least one booklet (hence 'balanced') and no booklet contains all

	LTT NAEP	Main NAEP (plus information on State NAEP)
	changes were made: ■ replacement of questions with outdated contexts ■ replacement of certain background questions ■ elimination of the 'I don't know' option category ■ use of single-subject-area booklets ■ discontinuation of audiotape pacing for mathematics (following a similar decision for reading made in 1984) ■ accommodations allowed. The question blocks for the 2004 trend assessments were assembled three to a booklet, together with a general background questionnaire that was common to all booklets. A partially balanced, incomplete block (pBIB) booklet design was used, which ensured that each block (and, therefore, question) was presented to a nationally representative sample of students and that each question was presented in various positions with respect to other questions. Since the booklets only contain question blocks from one	questions (hence 'incomplete') ■ each booklet is of roughly the same difficulty ■ during a particular session, a representative mix of booklets is distributed systematically across the student group (hence 'spiral') ■ there is no audiotape pacing ■ students respond to questions on a single subject area. By way of example, the most recent mathematics assessment: 'comprised 50 booklets at each grade. Each booklet contained two separately timed 25-minute sections of mathematics questions. The total numbers of test questions used in the 2003 mathematics assessment at grades 4 and 8 were181 and 197, respectively. Typically, a section, or block, contained approximately 16–20 questions, but there was considerable variation depending on the balance between multiple-choice and constructed-response questions.' (Taken from the

	LTT NAEP	Main NAEP (plus information on State NAEP)
	subject area, this is known as a 'focused' pBIB design.	2003 Mathematics report, p.4).
Student sampling model	It is hard to find a reasonably definitive answer to the question of how many students are involved; although a figure of around 5,000 students per LTT NAEP survey seems plausible. Samples sizes were larger for the 2004 surveys, at least partly because of the additional 'bridging studies' that were required. Nationally representative (probability) samples of participating schools are selected, from which students are chosen randomly.	It is hard to find a reasonably definitive answer to the question of how many students are involved, which is partly because sample sizes differ over time and across subjects (and are topped up for State and District analyses); although a figure of around 15,000 students per Main NAEP survey seems plausible (with samples topped up to around 3,000 per state for the State analyses). Nationally representative (probability) samples of participating schools are selected, from which students are chosen randomly. The sample sizes are calculated to ensure that at least 2,000 students respond to each assessment booklet. For the years in which State NAEP is run, reports are also presented at the national level, and the State NAEP samples are topped up with students from private schools and non-participating states. NAEP aims to be as inclusive as possible, and includes students with disabilities (SD) or classified as

	LTT NAEP	Main NAEP (plus information on State NAEP)
		limited-English proficient (LEP), allowing assessment accommodations where appropriate.
Assessment formats	Largely multiple-choice and short-answer constructed-response questions, including a small number of extended-answer constructed-response questions. Students are not required to explain their work.	Largely multiple-choice, short- and extended-answer constructed response questions. Students may be required to explain their work. Performance assessments may also be included. For example, the NAEP science framework says that: 'Innovative assessments in the United States and other countries use three major question types: performance exercises, open-ended paper-and-pencil exercises, and multiple-choice questions probing understanding of conceptual and reasoning skills… In performance exercises, students actually manipulate selected physical objects and try to solve a scientific problem about the objects… An extra period of time (20 or 30 minutes) may be necessary for students who have been assigned to perform complex tasks.'
Time	The assessment lasts for around 50 minutes per pupil, with around 45 minutes	The assessment lasts for around one hour per pupil, with (for example) around 25

	LTT NAEP	Main NAEP (plus information on State NAEP)
	answering questions from the subject domain and about five minutes on the general background questions.	minutes answering questions from two subject domain sections and about five minutes on each of the two general background sections (on home and school experiences related to achievement).
Administration	NAEP assessments are administered by trained staff, employed by the contractor for sampling and data collection, presently Westat, Inc. The school is asked to designate an in-school staff member as the school coordinator.	See left. (Each state also has a federally funded NAEP state coordinator who works with participating schools.)
Background data collected	Student and school characteristics.	Student and school characteristics. Four general sources provide context for NAEP results: ■ student questionnaires (background characteristics and teaching experiences) ■ teacher questionnaires (teacher training and classroom instruction) ■ school questionnaires (school characteristics and policies) ■ questionnaires on SD/LEP students

	LTT NAEP	Main NAEP (plus information on State NAEP)
		(students considered disabled or limited English proficient).
Scoring	Scoring guides (mark schemes) are developed by the contractor for design, analysis and reporting (presently ETS). Student responses are transferred from the contractor for sampling and data collection (presently Westat, Inc.) to the contractor for materials preparation, distribution, and scoring (presently Pearson). The marking process ensures that: ■ multiple-choice questions are machine-scored by optical-mark reflex scanning ■ constructed-response questions are scored by professional scoring personnel, using an on-screen marking system. All markers are fully trained and take qualifying tests. The quality of marking is regularly monitored. Results are fed into the NAEP database at the ETS.	See left.
Analysis	Analysis was originally conducted on a question-by-	Each subject domain is divided into sub-skills, purposes, or

	LTT NAEP	Main NAEP (plus information on State NAEP)
	question basis, reporting the percentages of students with correct answers (p-values), eg, 62% of 9-year-olds correctly identified that 'putting sand and salt together makes a mixture'. This was felt to be insufficiently useful, since the inferences that could be drawn lacked generalisability. With the transfer of responsibilities to ETS in 1983 and the use of IRT modelling, overall subject domain scores (and, hence, overall trend lines) could be produced. Overall average scale scores are now computed at the subject level, enabling the production of a single trend line for each survey. Average scale scores are also produced at various percentiles to illustrate potentially different trends for students at different levels of attainment. LTT NAEP also reports on the percentage of students reaching or exceeding each performance level, where the performance levels are defined in terms of five scale score points – 100, 150, 200, 250, 300, 350 – from a scale that ranges from 0 to 500. Students at these points are described in terms of the knowledge, skill and understanding that they	content domains and results are analysed for each of the principal subscales – mathematics, for example, has five subscales. Overall average scale scores are computed from the subscale scores, using statistical modelling based on IRT. The NAEP scales range from 0 to 300 or from 0 to 500, depending on the subject. Short-term trend lines can be produced, for subjects that are tested sufficiently frequently, on the basis of these overall average scale scores. Results are also reported in terms of the percentage of students within each achievement level. There are three achievement levels – basic (partial mastery), proficient (solid academic performance) and advanced (superior performance) – and performance descriptions illustrate what students deemed to be at each level ought to know, understand and be able to do. New achievement levels are established every decade or so, when the frameworks are revised. Data are weighted to ensure the representativeness of the

	LTT NAEP	Main NAEP (plus information on State NAEP)
	typically demonstrate. This process is known as 'scale anchoring'.	results.
Result reporting	Results are presented predominantly at the overall domain level, as single trend lines for each subject based on average scale scores. They are reported for the nation and for sub-groups defined by characteristics such as: ■ gender ■ race/ethnicity ■ parents' education level ■ type of school. Results not reported for individual students or schools. Questions from the LTT NAEP are generally not released.	See left. The reporting of achievement level results is an important aspect of Main NAEP. Although results are infrequently reported below the overall subject domain scale level, special reports are sometimes produced on specific subscales (for example U.S. DoE, 1999). Results tend to be released around six months after data collection for mathematics and reading, and around one year after data collection for other subjects. After each survey, around 25% of the test questions are made publicly available.
Reporting formats	Subset of right.	A variety of reporting formats are used, including: ■ **Report cards** – extended reports for policy makers that discuss results, design and administration for a single survey ■ **Highlights** – brief reports ■ **Snapshots** – very brief

	LTT NAEP	Main NAEP (plus information on State NAEP)
		state-level reports ■ **Update reports** – single issue reports for parents and the public ■ **Instructional reports** – assessment materials for educators ■ **State reports** – for state education executives, containing the results of a NAEP state assessment. ■ **Cross-state data compendia** – for state education executives and educational researchers ■ **Trend reports** – for educational researchers and policy analysts documenting long-term trends ■ **Focused reports** – for educational policy analysts and researchers, addressing policy issues ■ **Almanacs** – for researchers who wish to conduct secondary research ■ **Technical reports** – for educational researchers and psychometricians documenting procedures ■ **Demonstration booklets**

	LTT NAEP	Main NAEP (plus information on State NAEP)
		– with examples of questions ■ **NAEP questions tool** – on NCES website to provide examples of questions, responses and scoring guides ■ **NAEP data tool/explorer** – on NCES website to allow users to access customised analyses of data

Principal bibliography

There are very many published reports on NAEP stretching across the past five decades and written from a variety of perspectives, including: presentational reports on results; technical reports on the methodology; formal evaluations; social, political and educational accounts and critiques concerning the nature and impacts of the assessment; etc.

The most important resource for information on NAEP, including access to many reports and publications, is the official NCES website: http://nces.ed.gov/nationsreportcard/.

In addition to information taken from reports and summaries found on this site, the following texts were particularly useful in compiling this guide to NAEP.[3]

Beaton, A.E. (1990). Epilogue. In Beaton and Zwick (1990, below), pp. 165-168.

[3] The information presented in this guide has been extracted from multiple authoritative sources. Ultimately, though, since NAEP tends to be in a constant state of evolution, identifying the latest details of the programme can be problematic. In addition, even apparently authoritative texts occasionally contain contradictory claims. This guide should be read with these limitations in mind.

Beaton, A.E. and Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-1986 reading anomaly*. Education Testing Service. Princeton, NJ.

Duran, R.P. (2003). *Implications of electronic technology for the NAEP assessment.* US Department of Education. Institute of Education Sciences. Working Paper 2003-16.

Hambleton, R.K., et al. (2000). A response to 'setting reasonable and useful performance standards' in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

Johnson, E.G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29 (2), 95-110.

Jones, L.V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Education Researcher*, 25 (7), 15-22.

Jones, L.V. and Olkin, I. (2004). *The Nation's Report Card: evolution and perspectives.* Phi Delta Kappa Educational Foundation. Bloomington, IN.

Loomis, S.C. and Bourque, M.L. (2001). From tradition to innovation: standard setting on the National Assessment of Educational Progress. In G.J. Cizek (Ed.). *Setting performance standards: concepts, methods and perspectives.* Lawrence Erlbaum Associates, Inc.. Mahwah, NJ.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J (1999). *Grading the Nation's Report Card*. National Academy Press. Washington, DC.

Perie, M., Moran, R., and Lutkus, A.D. (2005). *NAEP 2004 Trends in academic progress: Three decades of student performance in reading and mathematics (NCES 2005–464)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

Shepard, L.A., Glaser, R., Linn, R.L. and Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. National Academy Press. Stanford.

US Department of Education (1999). *Estimation Skills, Mathematics-in-Context, and Advanced Skills in Mathematics*. Institute of Education Sciences. NCES 2000-451, by J. H. Mitchell, E. F. Hawkins, F. Stancavage, and J. A. Dossey. Washington, DC.

US Department of Education (2005). *The Nation's Report Card: An Introduction to The National Assessment of Educational Progress (NAEP)*. Institute of Education Sciences. NCES 2005–454 Revised. Washington, DC.

Vinovskis, M.A. (1998). *Overseeing the Nation's Report Card: the creation and evolution of the National Assessment Governing Board (NAGB)*. National Assessment Governing Board.

Insights from New Zealand's National Education Monitoring Project

Unlike monitoring systems in many countries, New Zealand's National Education Monitoring Project (NEMP) focuses on attainment in relation to a national curriculum. It is impressive in the degree to which it:

- assesses a full range of primary curriculum subject areas
- assesses as much as possible of the relevant domain for each subject area (including knowledge, skill and affect)
- employs a broad range of assessment formats, from traditional individual paper-and-pencil tests to novel group performance assessments (this is made possible, at some cost, by the involvement of trained administrators)
- embraces advances in information technology (IT) where appropriate – although only where appropriate – particularly through its use of video/computer presentation and recording.

NEMP is an example of a system designed to maximise the validity of inferences from results for formative purposes (formative for teachers, curriculum designers and policymakers rather than for pupils directly) and designed to ensure positive pedagogical impacts for teachers (through participation). However, these purposes and impacts are less prioritised by policymakers in England – since England's priorities focus more on accountability and performance measurement – and a national monitoring system designed to reflect these priorities might look quite different (in particular, it might be impossible to preserve the 'high validity' design characteristics of NEMP).

Introduction

The first NEMP surveys were administered in 1995, two years after the introduction of the New Zealand Curriculum Framework. Each survey aims to monitor attainment within, and attitudes towards, each of the primary subject areas of the framework, with an emphasis upon the use of authentic tasks. Monitoring occurs during the middle and end of primary schooling, that is, during year 4 (ages 8 to 9) and year 8 (ages 12 to 13).

Attainment and attitudes within each subject area are surveyed on a four-yearly basis, and trends between those years are examined. The aim is to provide information on how well overall national standards are being maintained, and on where improvements might be needed. As such, it allows successes to be celebrated and priorities for curriculum change and teacher development to be identified. The

primary goal of national monitoring is to contribute to the quality and improvement of student learning.

NEMP is run by a team based in the Educational Assessment Research Unit at the University of Otago (under Terry Crooks and Lester Flockton). It is guided by a National Advisory Committee, by Māori Reference Groups and Curriculum Advisory Panels.

NEMP is funded by the New Zealand Ministry of Education, to the tune of around NZD$2.5 million per year (representing less than 0.1 per cent of the primary/secondary education budget).[4] Almost half of this amount is spent on teachers, who are involved in task development, administration and marking. Beyond teacher involvement, NEMP is organised to foster a sense of teacher ownership.

NEMP surveys are 'low stakes', and neither school nor pupil results are released. Although participation is voluntary (for both pupils and schools), NEMP has excellent participation rates (partly because pupils generally find the tasks enjoyable and partly because teachers see the monitoring exercise as worthwhile).

Key features of the NEMP assessment model

As a relatively recent development, NEMP was based on lessons learned from a variety of international monitoring systems, particularly the National Assessment of Educational Progress (USA), the Assessment of Performance Unit (UK) and the Toronto Benchmarks project (Canada). It combines features of all of them, tailored to satisfy its particular purposes.

Assessed areas

NEMP aims to assess as much as possible of the New Zealand Curriculum Framework (1993), assessing each of its primary subject areas at four-yearly intervals:

1995, 1999, 2003	1996, 2000, 2004	1997, 2001, 2005	1998, 2002, 2006
science	language – reading, speaking	mathematics	language – writing, listening, reviewing
visual art	aspects of technology	social studies	health and physical education
information skills – graphs, tables, maps, charts, diagrams	music	information skills – library and research	

[4] These figures are taken from a 2001 report.

Within the survey for each subject area, NEMP assesses three facets of proficiency:

- knowledge (factual knowledge, conceptual understanding)
- skill (problem solving, physical, communication and interpersonal skills)
- affect (motivations, attitudes, values, dispositions)

Given its monitoring function (to maximise its resilience to curriculum change), it does not focus too tightly upon specific content features. It places more weight on conceptual understanding, skills and attitudes and less weight on knowledge of 'facts and figures'. It focuses specifically on 'important' learning outcomes and 'big pictures'.

Despite its monitoring function, NEMP is required to evolve in response to curriculum, pedagogy and assessment change over time (part of which occurs as a consequence of findings from previous surveys). Only short-term trends are explored, and primarily at the level of specific tasks.

Assessment tasks

NEMP assessment tasks are designed to:

- emphasise aspects of the curriculum deemed 'meaningful', that is, important to life
- be 'enjoyable', that is, motivating for pupils
- accommodate the full spectrum of ability.

With an emphasis on using the tasks most appropriate for eliciting the required assessment evidence – rather than those easiest to administer and mark – NEMP employs four task formats:

1. one-to-one teacher/pupil interviews (one pupil focusing on one task – oral presentation and response)
2. workstations (four pupils rotating individually around four tasks)
3. teams (four pupils collaborating on one task)
4. independent (one pupil focusing on one task – written presentation and response)

The majority of tasks are presented orally, on video or on computer (largely to avoid contamination of results by differences in reading ability); and the majority of

responses are provided orally or by demonstration (largely to avoid contamination of results by differences in writing ability).

For each subject area:

- about 35 per cent of tasks – known as 'trend tasks' – are repeated from the previous survey
- about 35 per cent of tasks – known as 'link tasks' – are new, but will be repeated in the following survey (when they will be known as 'trend tasks')
- about 30 per cent of tasks are new and will be used only once.

Given that link tasks will be repeated in the following survey, they are kept confidential, and their results are not discussed in detail. However, the remaining tasks (including trend tasks) are released and reported on. The methodology ensures that no task will be used more than twice, and there is an explicit intention that task design will evolve to keep pace with curriculum, assessment and pedagogical changes.

Sampling

Each year, around 1,440 pupils are selected for each of the two year groups. This represents around 2.5 per cent of the cohort (for each year). Pupils are selected through a two-stage process:

- a stratified quasi-random sample of 120 schools (excluding special schools)
- random samples of four pupils from each school for each of three task 'parcels' (12 pupils in total from each school).

This process is adjusted slightly to accommodate schools with small intakes, and there is some additional over-sampling of significant minority groups for sub-group analyses (although their results are used only for those sub-group analyses).

NEMP employs a 'large-block' task sampling model, for which:

- approximately four hours worth of performance tasks, for each of the three subject areas, are packaged into three task 'parcels'
- each of the 'parcels' includes approximately four hours worth of tasks, with all 'parcels' including tasks from each of the three subject areas
- each pupil attempts one of the three task 'parcels', and is assessed for approximately four hours, over a period of five days.

Since tasks are predominantly analysed discretely – there is no aggregation of results to domain or even sub-domain level – there is no need for any overlap in the tasks that pupils attempt (unlike more complex designs used in the US). Nor is there a requirement for the balance of tasks in any one year to correspond precisely to weightings within a content/process sampling framework.

Administration

All tasks are administered or supervised by experienced teachers, recruited and trained by the project staff. They are seconded for a period of six weeks, involving:

- one week of training at a central location
- five weeks, working in pairs, to conduct assessments for approximately 60 children.

Teacher-administrators work with no more than four pupils at any one time. They are required to give support with those features of task demand that lie beyond the focus of the assessment (reading/writing demands, where distinct from the assessment of reading/writing).

Recording and marking

The majority of responses to survey tasks are recorded on videotape or on computer, and all marking is undertaken centrally:

- senior tertiary students (typically trainee teachers) are employed to mark responses that can be scored objectively or with minimal judgement – they mark for approximately five hours per day for six weeks
- other responses are marked by teachers, recruited and trained by the project staff – they mark for either mornings or afternoons for one week.

The marking period lasts for approximately three months, typically requiring that around 3,000 hours of video-recorded performances and 60,000 pages of written responses be marked.

Analysis

Analysis and reporting is based on data for individual tasks and small clusters of highly related tasks; there is no attempt to create overall indices of attainment (either at the domain level or the sub-domain level).

Results are analysed predominantly at the national level, although a smaller number of analyses explore differences for sub-groups based on:

- pupil gender
- pupil ethnicity

- school region
- school community size
- school socioeconomic status
- school size
- school type.

There are three levels of analysis of task performance:

1. raw result analyses:
    - year 4 responses
    - year 8 responses
2. within-year comparisons:
    - year 4 responses versus year 8 responses
3. between-year (trend) comparisons:
    - year 4 (four years earlier) versus year 4 (four years later)
    - year 8 (four years earlier) versus year 8 (four years later).

Reporting

Before results are released, a national forum is convened to identify good news, concerns and suggestions for action. This results in the publication of the *Forum Comment*, which is sent to all teachers in New Zealand.

Full technical reports are produced on each subject area tested each year (which are published on the NEMP website). Finally, 'access tasks' are made available for classroom use (these need to be purchased).

Insights from the NEMP model

Unlike monitoring systems in many countries, NEMP is focused on attainment in relation to a national curriculum. However, given its monitoring purpose, it still emphasises the broader aspects of attainment: those central to the subject areas and least likely to change in relevance over time.

It is impressive in the degree to which it:

- assesses a full range of primary curriculum subject areas

- assesses as much as possible of the relevant domain for each subject area (including knowledge, skill and affect)
- employs a broad range of assessment formats, from traditional individual paper-and-pencil tests to novel group performance assessments (this is made possible, at some cost, by the involvement of trained administrators)
- embraces advances in IT where appropriate – although only where appropriate – particularly through its use of video/computer presentation and recording.

These design features support the following valued outcomes of monitoring:

- lessons for teacher development and the curriculum, at some level of detail (specific to narrowly defined dimensions of competence, as embodied in the specific tasks employed)
- positive 'washback' for the teachers who are involved in the project as developers, administrators and markers.

On one hand, NEMP is an example of a system designed to maximise the validity of inferences from results for formative purposes (formative for teachers, curriculum designers and policymakers, rather than for pupils directly) and designed to ensure positive pedagogical impacts for teachers (through participation).

On the other hand, though, NEMP is not an example of a system designed to maximise the validity of inferences from results for a variety of non-formative purposes, for example long-term monitoring or performance target accountability. The principle of evolving assessment instruments over time presents a major challenge to the long-term monitoring purpose, while the lack of overall aggregate results presents a major challenge to the accountability purpose. A decision to prioritise the long-term monitoring purpose would recommend non-release of items, which would restrict the reporting considerably and threaten the formative purpose. Analogously, a decision to prioritise the accountability purpose would recommend the use of complex statistical techniques (such as item response modelling), which would restrict the range of assessment tasks and areas and would threaten the validity of results more generally.

In short, while NEMP represents a very attractive national monitoring system, and one that operates in the context of a national curriculum akin to that in England, its viability is intimately related to the uses and impacts that policymakers presently prioritise. These priorities would not necessarily be shared by policymakers in England.

It remains to be determined whether even policymakers in New Zealand will remain satisfied with NEMP as it is presently conceived. NEMP has been in existence for only a short period of time; that it lacks power in relation to monitoring even medium-

term trends, and in relation to even relatively narrow sub-domains, may present more of a political challenge as time goes on. In fact, it seems that observers already increasingly wish to draw broader inferences about attainment trends, at levels beyond the individual task, to give a more aggregated picture of performance patterns (Terry Crooks, personal communication, 29 August 2005). It will be interesting to see the extent to which the present design can support this intention, or whether there will be pressure to move away from the present design.

It also remains to be seen the extent to which NEMP is felt to generate genuinely important lessons for teacher development (its primary purpose). For example, in a media release, dated 28 July 2004, from the Ministry of Education (entitled *NEMP study shows primary students making progress*), we read the following observation: 'Students in both year 4 and year 8 were least successful in providing titles and appropriate labels for axes and values, seeming to believe that it was sufficient to display the data.' This seems to offer an obvious lesson for teachers, but not necessarily one that is particularly 'deep'; the error to which it relates seems unlikely to signify a major conceptual deficit. Were all teachers to focus on this error, it would be straightforward to elevate performance on similar tasks in the future, but would this tell us much about whether students made a significant advance in the relevant sub-domain of information skills?[5] Maybe; maybe not. Again, the key issue here is the extent to which lessons from performance on narrowly defined tasks are seen to generate genuinely important lessons for teacher development. Only time will tell.

Bibliography

This chapter was based upon information from the following sources:

Flockton, L. (1999). *School-wide assessment: National Education Monitoring Project. New Zealand Council for Educational Research*. Wellington, New Zealand.

Crooks, T. (2002). *Design and implementation of a National Assessment Programme: New Zealand's National Education Monitoring Project (NEMP)*. Paper presented to the Annual Conference of the Canadian Society for the Study of Education (CSSE). Toronto, 25-28 May.

Documents on the website: http://nemp.otago.ac.nz/

Thanks also to Professor Terry Crooks who provided additional information.

[5] Even if they had, might this be at the expense of an element of attainment that now received less attention? The task-by-task analysis, which resists overall aggregation, can make it hard to detect this kind of 'swings and roundabouts' effect.

Developing a system for monitoring national educational attainment trends: purposes and decisions

This report intends to:

- emphasise the importance of deciding – in advance of any decisions concerning system design – the primary purpose for which results will be used
- illustrate the range of system design decisions that will need to be made.

There are many possible purposes for which results from a national attainment monitoring system could be used, including (among others):

1. research
2. target tracking
3. system warning
4. curriculum evaluation
5. teacher development.

Each of these purposes would recommend that the monitoring system should be designed somewhat differently. This highlights the importance of deciding, in advance, the primary purpose for which results will be used. Once this decision has been made, further system design decisions can be made. These will concern the following features (among others):

1. management structure
2. domains assessed
3. assessment frameworks
4. survey frequency
5. assessment tasks and formats
6. student groups
7. question sampling and presentation model
8. student sampling model
9. administration and response recording

10. background data collected

11. scoring

12. analysis and reporting

13. reporting formats

14. evaluation.

Introduction

This report highlights issues that ought to be considered before developing a system for monitoring national educational attainment trends.

Although focused on the national level, this kind of system could conceivably extend to explore trends at a regional level. However, it would not extend to explore trends at a local, school, class or pupil level (for which alternative assessment arrangements would be required).

Although focused on educational attainment, this kind of system is unable to distinguish between attainment resulting from learning that occurs in school and attainment resulting from learning that occurs outside of school (for example at home, within clubs, private tuition, during everyday life). It concerns educational attainment in the sense of focusing on the forms of knowledge, skill and understanding that are typically taught to all students, having been deemed an important part of preparation for everyday life.

The following discussion is couched mainly in terms of the assessment of discrete subject domains, although the surveys do not necessarily need to be structured in this way.

Common design features

A number of design features will likely be common to most national attainment monitoring systems. These include:

1. assessment of those aspects of each subject domain considered to be **least likely to change in social/educational significance** over time (since the validity of inferences from attainment trends are seriously threatened when aspects of subject domains change in significance over time)[6]

[6] By way of example, consider the present-day significance of the slide-rule, terms like ‘wheelwright’, and computer punch cards, among others.

2. assessment of **different students on different blocks of tasks** (to sample each subject domain thoroughly using a full range of assessment formats)

3. assessment of a **representative sample of each cohort**, rather than entire cohorts (to ensure financial and administrative feasibility)

4. use of the **same frameworks, tasks and procedures** over time (since change in any of these variables renders the system vulnerable to error)

5. **administration under controlled conditions by trained staff** (since the administration procedures are often not straightforward, even for written tasks, and to ensure that the security of materials is not breached)

6. **scoring of responses undertaken externally** (to ensure consistent accuracy of marking)

7. **not releasing the specific monitoring tasks** for public scrutiny (since if used in class by teachers – either intentionally or unintentionally – this would invalidate inferences from performance trends)

8. ensuring that **results have low stakes** for schools/pupils and **not reporting** those results (so that teachers have little incentive to drill students in techniques for performing well in the kind of tasks upon which the monitoring is based – where such techniques do not actually implant robust understanding – and so that teachers and students have little incentive to breach security by revealing specific tasks)

9. **analysis of performance by different subgroups** of the population (to explore the generalisability of conclusions drawn at the national level).

There are numerous other design features that differ between monitoring systems, depending upon the primary purpose for which results are intended to be used.

Alternative monitoring purposes

A variety of possible monitoring purposes exist, including (but not limited to):

1. **research** – to identify social/educational/political (among other) factors associated with improvement in attainment over time

2. **target tracking** – to investigate progress towards national attainment targets

3. **system warning** – to notify the nation of potential educational problems (especially of unanticipated attainment impacts that might have occurred as the result of policy change)

4. **curriculum evaluation** – to identify priority areas for development

5. **teacher development** – to improve curriculum and assessment literacy among teachers.

It would not be feasible to design a monitoring system perfectly suited for all monitoring purposes simultaneously. In fact, a system designed to satisfy one monitoring purpose might be entirely unsatisfactory for another monitoring purpose. For this reason, the policy maker's **primary purpose** needs to be made explicit *in advance of any system design decisions*. If there is an aspiration for the system to serve more than one purpose, then **trade-offs** in system design may need to be made; these may render the system less than perfectly suited for either purpose. The extent of compromise will need to be evaluated in advance to determine whether results are likely to be **sufficiently fit** for those primary purposes (and, if not, then the policymaker's aspiration may need to be revisited).

Different design features

The importance of tailoring the design of a monitoring system to the specific purpose for which results will be used can be illustrated by highlighting features that would be particularly significant for each of the purposes listed above.

Research

If a monitoring system is to produce results that can be used for research purposes, then it will be important to collect background information on characteristics of participating students that can be statistically modelled against their performances. These might include, for example, their:

- region's characteristics (for example number of schools, population density)
- school's characteristics (for example number of pupils with free school meals, school size, management policies)
- family's characteristics (for example number of siblings, parental socioeconomic status)
- own characteristics (for example season of birth, IQ, handedness).

The precise balance of background characteristics chosen would reflect the particular research interests prioritised and hypotheses concerning potentially causal factors.

During the early years of the UK's Assessment of Performance Unit (APU) monitoring system, there was considerable debate over whether it should fulfil a research function. Although mooted, detailed research studies never materialised, limiting the usefulness of conclusions.

Target tracking and curriculum evaluation

Inferences from trends might be drawn at a variety of levels, for example:

1. [24 x 34 =]
2. [67 x 84 =] [45 x 21 =] etc
3. [574 x 28 =] [456 x 34,784 =] etc
4. [454 / 789 =] [(67 - 90) / (78 + 23) =] etc
5. [79.57 / 19.7908 =] [(34.6 / 5) x (78.5 / 3) =] etc

The first level (above) is the task level. If average performance improved from one survey to the next, this would support the inference that the cohort was better at the specific task: [24 x 34 =]. Were performance to improve generally across tasks at the second level (above) this might support the inference that the cohort was better at two-digit whole number multiplication tasks. If performance were to improve generally across tasks at the fifth level (above) this might support the inference that the cohort was better at multi-digit variable-integer multi-basic-function numerical tasks. Of course, the list could be extended down until the level of inference was so broad that it encompassed the entire subject domain; a domain that might be defined as 'essential everyday mathematics'.

Inferences from trends are most straightforward when they relate to performance on specific tasks. If task performance improves, then it has improved; that is all there is to say. However, the inference only extends to that specific task and – by itself – is not of a great deal of interest.

It becomes more interesting when inferences can be drawn across groups of similar tasks; for example, two-digit whole number multiplication tasks. Here, the improvement would describe improved attainment of a very specific skill. However, the potential drawback is that, although performance might improve across the majority of tasks, there might still be a substantial minority of these tasks for which performance deteriorated; so the inference would be less clear-cut.

By extension, when inferences are drawn at the subject domain level, underlying patterns of improvement and deterioration are lost (even potentially important ones, such as an increase in multiplication skill being 'balanced out' by a decrease in long division skill and appearing as no change in overall attainment).

The purposes of curriculum evaluation and target tracking typically fall at opposite ends of the continuum in relation to the specificity of inferences from trend results. To explore the strengths and weaknesses of curriculum delivery, a relatively fine-grained analysis of trend results is necessary. On the other hand, it would be entirely

unwieldy to establish national performance targets at a fine-grained level and targets are usually set at the subject domain level.

The distinction is not at all trivial. To produce results at the subject level, complex statistical techniques are required. These, in turn, necessitate that performance across different tasks tends to conform to particular statistical patterns. If certain tasks tend not to conform well to such patterns, then a decision may have to be made to exclude them, even when they represent very important areas of the curriculum.

New Zealand's National Education Monitoring Project (NEMP) functions more towards the curriculum evaluation end of the continuum. It reports at a fairly low level but is able to embrace a very full range of assessment objectives within each subject domain and uses a very full range of assessment formats (since it does not aggregate to a subject level using complex statistical modelling of task responses). By contrast, the US's National Assessment of Educational Progress (NAEP) functions more towards the target tracking end of the continuum. It reports primarily at the overall subject domain level (using complex statistical modelling of task responses), but is quite constrained in terms of assessment objectives and formats.

The restriction of range of assessment objectives and formats can be a major problem for systems that report primarily at the overall subject domain level, since those aspects that are not assessed may well be very important. Inferences about the entire domain – from performance on a restricted range – may be misleading. This can lead to ineffective policy making.

System warning

A system designed to satisfy a system warning purpose will likely need to operate, somewhat in the policy background, over a period of decades rather than years. This will impose two major requirements to make it as resistant as possible to the challenge of comparability of standards over time. It must embrace aspects of core subject domains that:

1. represent their essence

2. are least likely to change in significance.

These are related, although slightly different, requirements. The first insists that the monitoring system be capable of detecting change across each and every aspect of a subject domain that is deemed central to it. There is little point in including aspects of knowledge, skill and understanding that are peripheral (those aspects for which it would not matter much if their prevalence within the population did change significantly over time). The second insists that the monitoring system can only provide robust trend lines as long as the aspects of knowledge, skill and understanding that are assessed in the baseline year continue to have similar real-

world significance in future years. Were they to become less relevant as time went by, the apparent decline in attainment over time would have little significance; it would not necessarily sound any useful warning.

Although it is not possible to predict the future with accuracy, the second requirement is likely to steer the system designer towards aspects of skill and understanding and away from aspects of knowledge. To some extent, this might distinguish a system designed for system warning from one designed for curriculum evaluation (although this is likely to be a matter of degree). The system warning model, like the curriculum evaluation model, is likely to report at a fairly fine-grained level.

Teacher development

A system designed to improve curriculum and assessment literacy among teachers must, obviously, be capable of engaging teacher interest and motivation. There are different ways of doing so.

New Zealand's NEMP has teacher development as an explicit aim. Each year, it invites a large group of practising teachers to participate in the national monitoring exercise, either as administrators or as markers (giving priority to teachers who have not participated previously). Administrators attend a week-long training course – which trains them to administer complex performance assessments with individual students and with groups of students – after which they spend five weeks actually assessing students. The administrators work in pairs, which presumably enhances the learning experience.

Since NEMP is more focused on short-term than long-term trends, it has more opportunity to release a substantial proportion of tasks to schools, which again supports the teacher development function.

The variety of possible design decisions

Different monitoring systems may differ radically in terms of a variety of possible design decisions. The precise reason for each design decision will depend (in theory) upon the primary purpose for which results will be used and (in practice) upon more pragmatic features of the context in which the system will operate that act as **constraints** upon the ideal design. In designing a monitoring system, it must constantly be borne in mind whether particular constraints, compromises and trade-offs will render results from the monitoring system insufficiently fit-for-purpose. The following list illustrates the range of decisions that will need to be made, once the primary purpose has been decided.

It is essential to recognise that the decisions will have to be made through **interplay of technical and political concerns**; political concerns will steer the specification of purposes while technical concerns will steer the choice of system features to best satisfy those purposes. The technical issues involved in making each decision are

particularly complex, owing to inter-dependencies between them. For example, the apparently straightforward question of how many students should be sampled for each survey is very far from straightforward, depending on a host of other decisions (such as how many tasks are used, level of analysis of results, type of statistical model used, approach to task sampling and length of time each student is assessed).

Design specifications

Management structure

- Where should responsibility for overall programme management reside?
- How will commissioning of experts (for example assessment, research and statistics) be managed?
- How will stakeholder perspectives (for example professional bodies, teachers, employers and subject communities) be managed?
- How should the division of responsibilities for survey design, administration, scoring, analysis, reporting (among others) be allocated (entire programme contracted out to a single organisation; different organisations contracted to run different subject surveys; different organisations contracted to run different aspects of the surveys for all subjects)?
- What overarching groups need to be established (for example steering, advisory and working), and what should be the limit of their responsibilities?
- How can long-term staffing be ensured to enable the kind of collective memory important for long-term monitoring systems?

Domains assessed

- Should the surveys be divided up along traditional subject lines (for example reading, writing and science), subject-related lines (for example literacy and numeracy) or along entirely different lines (for example problem-solving, researching and communicating)?
- If traditional subject lines are chosen, should all curriculum areas be included or only some (if the latter, then which)?
- Even if traditional subject lines are chosen, should the domains explicitly include aspects that may lie beyond what is traditionally studied at school?
- Who should decide the definition of assessed domains and frameworks (for example subject communities, policymakers and panels of citizens)?

Assessment frameworks

- To what extent should the assessment frameworks – the detailed specifications for each domain – focus on traditional components of knowledge, skill and understanding, as studied at school?
- What balance of content (knowledge) and process (skill) should be reflected?
- To what extent should broader constructs (such as problem solving, researching and communicating) be included, either within specific domains or as cross-domain themes?
- To what extent should the system attempt to monitor attitudes?

Survey frequency

- How frequently should each survey be run (for example annually, bi-annually, four-yearly)?

Assessment tasks and formats

- To what extent should each survey be based upon traditional formats (for example written responses to written tasks) or more novel formats (for example computer-based responses to computer-based tasks, performance-based and video-recorded responses to orally presented tasks or group work)?
- How open-ended may responses be, and should students be required to explain their work?
- Should preference be given to task formats that are easy/cheap to administer and score, or to task formats that will maximise the validity of inferences from results and secure (where intended) positive washback?

Student groups

- Should sampling be on the basis of age group (for example nine-year-olds) or school cohort (for example year 7)?
- How many and which age groups / cohorts should be the focus of each survey?
- Should young adults beyond school leaving age be included?
- Should only state school students be included, all school children, or all children (including even home educated)?
- Should students with learning difficulties, severe sensory handicaps, English as an additional language, or other circumstances be accommodated?

- Should, where possible, participation be made compulsory?

Question sampling and presentation model

- How many tasks should be used to assess each domain?
- How many sub-tests should be formed to spread these tasks across students?
- To what extent should tasks overlap across sub-tests?
- Should sub-tests include tasks from single domains or multiple ones?
- To what extent is computer adaptive testing desirable or feasible?

Student sampling model

- Should students be sampled by school (whereby all students in a sampled school), or should a two-stage model be applied (whereby schools are sampled first and then pupils are sampled from within those schools)? This will be more complex if including home educated children.
- Should the sampling be entirely random, or should representativeness be ensured through alternative mechanisms?
- How should pupils be sampled from small schools?
- How should sufficient sampling of small subgroups be achieved?
- How should non-participation of sampled schools/pupils be accommodated?
- How can the principle of 'light sampling' be achieved if many surveys are running simultaneously?

Administration and response recording

- How should administrators be appointed?
- How much training should administrators be given?
- How many administrators will be needed?
- To what extent should tasks be presented and/or recorded using personal computers (PCs)?
- To what extent should administrators assist students in understanding task demands (beyond the standard instructions)?
- To what extent should performances be recorded or evaluated directly?

- Should tasks be administered individually or in groups?
- For how long should each student be assessed?

Background data collected

- What kind of background data needs to be collected, either for research purposes or simply for subgroup analysis/reporting (for example pupil gender, pupil ethnicity, school region, school community size, school socioeconomic status, school size and school type)?
- What is the best way to collect background information reliably, sensitively and confidentially (for instance student questionnaires, teacher questionnaires or school questionnaires)?

Scoring

- How should scorers be appointed?
- How much training should scorers be given?
- How many scorers will be needed?
- To what extent can/should scoring be automated?
- What quality assurance/control mechanisms should be put in place?

Analysis and reporting

- Should analysis and reporting occur at task level, sub-domain level and/or domain level?
- If reporting at sub-domain or domain level, what approach to statistical modelling should be adopted (for example generalisability modelling, item response modelling)?
- What steps can be taken to ensure that analysis and reporting – which will not happen until the programme has been specified and in operation for a number of years – remains directly related to the primary purpose (and not to any other purposes that might be mooted along the way)?
- What approaches can be taken to modelling error variance (specifying parameters according to which the significance of change can be judged) and how can this be communicated?

- Should results be presented in terms of the percentage of students able to do x, y or z, in terms of what students at a particular percentile can do, or in some other format?
- What kind of questions will it not be possible to answer using the chosen monitoring system?

Reporting formats

- For whom should results be interpreted and presented?
- How can it be ensured that targeted user-groups receive, process and understand the results?
- How best can performance be illustrated when tasks need to be kept secure?
- How should tasks be released (how many and when)?

Evaluation

- How should evaluation be built into the programme?
- To what extent should evaluation be routine and scheduled or targeted on apparent anomalies?

Characteristics of a system for monitoring progress towards public service agreement targets for education

The aim of this report is to illustrate, and to explain, features of assessment that should characterise an effective system for monitoring progress towards public service agreement targets for education (expressed in terms of the level of attainment of a national cohort in a range of subject areas).

Purposes and design characteristics

Results from national monitoring systems might serve any of a variety of purposes, each with different implications for the design of those systems. For example, a system intended to monitor curriculum strengths and weaknesses would not need to aggregate results across the distinct sub-domains of a subject area; whereas a system intended to monitor progress towards national public service agreement (PSA) targets would.

The PSA target system would need, for each cohort studied, to produce results at an overall level for each subject domain studied (for example science). In contrast, the curriculum strengths and weaknesses system would need to present results at lower levels (for example the sub-domains of nutrition, circulation, movement, growth and reproduction, and health; within the sub-domain of humans and other animals; within the sub-domain of life processes and living things; within the domain of science). The latter would enable policymakers to see micro-trends across sub-domains of subjects (for example a general increase in understanding of nutrition against a general decrease in understanding of circulation), but not macro-trends across subjects. The former would enable policymakers to see macro-trends across subjects (for example a general increase in understanding of science), but not micro-trends across sub-domains of subjects.

The point of this extended example is to emphasise a fundamental point: **the decision over what kind of inference you want to draw from results – the choice of monitoring purpose – will directly impact upon how the system ought to be designed.** A system could be designed to report either at the micro-level, macro-level, or conceivably at both levels. However, to report at both levels would require the most technically complex of designs, with the most significant commitment of resources, and would be the most likely to be susceptible to assessment error. **Susceptibility to assessment error is a very serious risk when monitoring trends, so a system should be only as complex as absolutely necessary.** This means deciding carefully in advance exactly what kind of inferences need to be drawn from results and what kind of inferences need not, and planning the system accordingly.

Without developing the argument for prioritising this purpose in any detail, the following sections will illustrate, and explain, features of assessment that should characterise an effective system for **monitoring progress towards national PSA targets.** This would necessitate the computation of aggregate scores to represent attainment across subject domains but would not require the computation of scores to represent attainment within sub-domains, nor support inferences below the level of the subject as a whole. It would necessitate a fairly high level of technical complexity and resource commitment.

Monitoring progress towards national PSA targets for education

There are general characteristics that any assessment designed for monitoring educational attainment ought to exhibit, and specific ones for an assessment designed for monitoring progress towards national PSA targets for education (see table 1, at end of report).

The most important general characteristic is that it should assess only the **core elements** of each subject domain – those central to the subject and least likely to change in social or educational significance over time – and that the **core framework**, which specifies aspects of content knowledge and process skill, should remain unchanged over the period of monitoring. In short, the system should aim to measure the *same* construct from one year to the next, and that construct should be as *relevant* at the beginning of the period of monitoring as at the end.

However, it should be stressed from the outset that there are substantial technical and practical challenges of both specific and general nature:

- the technology of measuring trends is challenged principally by the tension between a conceptual requirement to ensure that measurement procedures remain constant over extended periods in time and a practical inevitability that **measurement procedures will change** (even very slight, apparently trivial, changes can corrupt trend lines)
- this is exacerbated when measuring trends in educational attainment because, even when measurement procedures remain constant over time, **the measured construct may change** (for example items that are a good index of 'spelling ability' in 1960 may no longer remain so in 2010)
- this is exacerbated further when measuring progress towards targets, since (and this is a not very pretty expression of Goodhart's law):
    - what is actually measured is merely an index of (and tends to be narrower than) what is intended to be measured
    - action will focus upon what is actually measured rather than what is intended to be measured

- at best, **increases in what is actually measured may overestimate increases in what is intended to be measured**
- at worst, **monitoring may corrupt the system being monitored** when action that increases what is actually measured is too far removed from action that increases what is intended to be measured.

Unless specific features are built into the design of a national monitoring system, these challenges will defeat the monitoring enterprise. However, no system – however well designed – could be entirely immune to these kinds of challenges.

Table 1 presents and explains design characteristics for an assessment system intended to monitor progress towards national PSA targets for education.

From table 1, **the most important decision to be made is whether to compromise on the length of the assessment.** First principles would require an assessment long enough fully to represent the domain (of content knowledge and process skill) defined by the core framework for each subject area, using an appropriate range of assessment formats (task types). Generally speaking, this is assumed to require an extended battery of tasks that, in total, might occupy (something in the range of) 15 to 40 hours (or considerably more).[7] To accommodate this much assessment, complex designs for sampling, administration and analysis must be constructed, so that different students attempt different combinations of tasks and results are 'stitched together' statistically (a task sampling model). Unfortunately, this is resource intensive, and its complexity renders it vulnerable to assessment error.

The alternative would be for all students to be administered a single 'indicator' or 'reference' test – the same test each year – and for that test to sample the core framework as well as possible, using only a limited number of assessment formats (a single test model). In a national monitoring context, where there is no external incentive for students to perform optimally, this test would have to be short (less than an hour).

Although the single test model has been used in the past for monitoring purposes, it is generally no longer considered adequate (particularly given developments in statistical modelling over the past half-century). It is dangerous for many reasons, including the threat to validity from a breach of test security (even assuming low stakes). If the possibility existed that students might have been coached using the

[7] The precise time requirement would probably differ by subject and, to sample as thoroughly as possible, would probably take very much longer than illustrated (the illustrated range reflects operational decisions that have been made for a number of systems in the past).

specific kinds of task that appeared in the monitoring test, then it would be impossible to trust results from the test as a valid index of what was supposed to be measured.

Even appreciating the additional risks and costs attributable to the added complexity of the task sampling model, it should still be recommended. Not to adopt the task sampling model would be to fall short of international standards for national monitoring systems.[8]

Finally, it should be recognised that **PSA targets** – based on results from the kind of monitoring system described in this report – **would have to be defined in terms of the narrow construct of educational attainment given by each core framework.** The trend lines would not, for instance, measure the broader constructs of educational attainment given by national curriculum programmes of study. As such, genuine improvements in educational standards that did not impact upon elements of the core framework would not register in results from the monitoring surveys.[9] In short, trend lines from national monitoring surveys and national curriculum tests would not, and should not, necessarily look the same.

[8] See, for example, the commentary on national assessments on the World Bank website: http://www1.worldbank.org/education/exams/nature.asp

[9] A word of clarification is required here. National curriculum tests are often criticised for failing to assess the full range of learning outcomes from their frameworks (their programmes of study). National monitoring surveys, in contrast, should aim to assess the full range of learning outcomes from their frameworks. However, their frameworks are narrowly defined (restricted to the core elements of a subject domain). Even though national curriculum tests might not assess the full range of learning outcomes from their programmes of study, they might still assess a broader range of learning outcomes than the national monitoring tests.

Table 1: Design features for an assessment system intended to monitor progress towards PSA targets

Design principle	Design implication	Rationale
System designers must begin by making explicit – for present and future audiences – the kind of inferences the system will support (and the kind of inferences the system will not support). This should define the purpose of the monitoring system.	The system 'remit' should be produced in advance of any subsequent development work. It should be published.	Compelling evidence shows that failing to do so will result in the system failing to satisfy stakeholder expectations (in particular, from England's experience of the Assessment of Performance Unit, which failed to monitor attainment standards effectively).
System designers need to take a long-term perspective in every decision they make, regarding the following:		Any monitoring system will fail unless, from the outset, longevity is elevated as its central design principle.
■ management and development teams and procedures.	Long-term funding needs to be committed from the outset. Management teams need to be located within secure units (secure, that is, from an employment perspective). The retention of key staff needs to be a priority.	The development of a national monitoring system is a highly specialised enterprise. The management and development teams would be likely to have to develop entirely new skills and knowledge bases and, as such, would become national experts. The accumulated expertise of such teams would be unique, from a national perspective; and retention of key

Design principle	Design implication	Rationale
		staff would be essential. Ultimately, although procedures would be written down, there is no substitute for past experience in guiding future practice. This is never truer than when complex procedures need to be replicated from one year to the next.
■ assessment frameworks.	Assessment frameworks should comprise only the core elements of attainment in each subject domain, those deemed most important for all students to acquire, to enable effective participation in everyday life. Students will acquire these competencies both through formal education and everyday experiences beyond school. Assessment frameworks should also focus only on those elements of attainment in each subject domain that appear least likely to change in social and educational significance over time. More generally, assessment frameworks must remain identical over time.	The first priority is that the assessed content knowledge and process skills should represent the essence of a subject domain, so that inferences from trend lines will say something important. The second priority is that the assessed content knowledge and process skills of the framework should (in all probability) be considered just as important in 10 to 20 years time as they are today. This requirement enables the measurement of change. (More generally, it is not possible to measure change in the amount of something from one decade to the next, unless that 'something' remains constant.

Design principle	Design implication	Rationale
	For certain subject domains, such as information and communication technology (ICT), it may be very hard, if not impossible, to identify core elements of content knowledge and process skill unlikely to change in social and educational significance over time. If so, then monitoring should be avoided entirely.	If it changes, then you will simply have measured an amount of something at point 1 and an amount of something else at point 2. Therefore, not only do you need to assess the same framework over time, the framework needs to retain its relevance.)
▪ assessment tasks, mark schemes, administrative procedures.	Assessment tasks and mark schemes should, when possible, remain identical over time. Administrative procedures should only be changed when unavoidable, and the impact of any change should be modelled through a 'bridging study'. Since assessment tasks remain identical over time, their security needs to be maintained over time as well. Administration should occur under controlled conditions using trained staff. The tasks will similarly need to be marked under controlled conditions, by trained	Again, constancy is a basic requirement for the measurement of change. Even when performance on the assessment has no direct consequence for schools or students, temptation likely will still exist for teachers to want to use the tasks within their lessons (after all, the tasks will have been designed to assess core competencies very effectively). This presents a threat to test security even under conditions of low stakes. (This threat is exacerbated when the core framework is quite narrow.) Test security needs to be ensured through a variety of mechanisms, one of

Design principle	Design implication	Rationale
	staff.	which is administration by trained staff (rather than by teachers or invigilators). Trained staff will also ensure the constancy of application of procedures. Administration of tasks and the recording of responses might be facilitated by the use of ICT. However, given the propensity for rapid obsolescence of ICT platforms – and the necessity of keeping administration procedures identical over time – this possibility should be approached cautiously.
▪ research, sampling and statistical designs.	Research, sampling and statistical designs should remain identical over time. Moreover, they should be designed to accommodate foreseeable (unavoidable) changes. This might, for example, require the use of statistical designs based on item response modelling, to enable the identification and replacement of tasks that fail.	Again, constancy is a basic requirement for the measurement of change.

Design principle	Design implication	Rationale
▪ reporting practices.	Reporting practices should remain identical over time.	Again, constancy is a basic requirement for the measurement of change.
▪ quality assurance.	Evaluation teams and remits need to be established from the outset, to design and implement procedures for ensuring that inevitable measurement challenges are overcome.	Again, constancy is a basic requirement for the measurement of change.
Assessment tasks should sample from the full subject domain (of content knowledge and process skills) as defined by the core framework, and should sample from an appropriate range of assessment formats.	This will require a large battery of assessment tasks, comprising many hours in total. Since it is not feasible to assess individual students for extended periods, the battery of assessment tasks will have to be split up and distributed across students according to a complex sampling design. The appropriate range of assessment formats might well include selected response (for example multiple-choice questions); constructed response (for example written short-answer questions, written essay questions, vivas); performance (for example orals, practical tasks).	Thorough 'construct representation' is a basic measurement requirement that ensures what is measured corresponds as closely as possible to what is intended to be measured. This is of particular importance for a system that operates in the context of performance targets (as explained earlier). [The alternative would be to have a single test that assesses a limited coverage of the domain using a limited range of assessment formats. However, this would be a dangerous model upon which to base a monitoring system, since genuine changes in aspects that were not sampled would not be measured. Paradoxically, the aspects that tend not to

Design principle	Design implication	Rationale
	Results from different students on different tasks will have to be aggregated together using complex statistical modelling procedures (for example item response modelling).	be sampled (for practical reasons) – the more complex dimensions of content knowledge and process skill – are often considered the most important. The danger of adopting the single test model would be compounded during national monitoring, since the length of the test would need to be kept to an absolute minimum to ensure that students remained motivated throughout it. The accuracy of inferences from results relies on students performing as well as they ought to. Since there are neither external rewards nor sanctions for performing optimally, the success of the monitoring is dependent upon the cooperation of students. Thus, single monitoring tests would likely be shorter even than traditional examinations (less than one hour). The danger of adopting the single test model would be compounded even further if, in addition to national (survey) monitoring, separate (national curriculum) tests were also administered for purposes

Design principle	Design implication	Rationale
		like school comparison. The decision to adopt a single test model for monitoring purposes would very likely mean that the kinds of tasks (in the survey tests) used for monitoring would be a 'subset' of those (in the national curriculum tests) used for comparison. Consequently, the impact of coaching for the high stakes comparison tests would generalise to performance on the low stakes monitoring tests, reducing the validity of inferences from trends.]
There should be no direct consequences of good or bad performance for those assessed (whether schools, students or local authorities).	Neither individual nor aggregate results should be reported back to participants. Results should only be reported for the nation, and for large subgroups (for example breaking results down by gender, ethnic background, region, and other ways).	This is important to minimise the impact of monitoring on pedagogy (to ensure that the results can be taken to mean what we require them to mean). For this reason, but more directly, the lack of direct consequences should mean a reduction of any temptation to breach test security.
It is advisable **not** to assess the following:		
▪ all students within a	A sampling framework should be devised that represents the distribution of various	There are many reasons for not testing all students. The most obvious is a reduction

Design principle	Design implication	Rationale
particular year group.	subgroups accurately. The sampling will likely include all schools (at least over time) but will involve selecting only a small number of students from those schools (probably at random). When drawing inferences about the population as a whole, students who do not attend standard state-maintained schools will also need to be sampled (for example children in independent and special schools, home-educated children or students who leave school early).	in financial cost. Where administration procedures are as complex as national monitoring tends to require, cost is a fundamental consideration: the administrative burden due to complex sampling, multiple task production and scoring, complex tasks and other factors would require massive funding if carried out for the population (rather than for a sample). Sufficient accuracy can be achieved on far smaller samples, even for subgroup analyses below the cohort level. Equally important, though, is the reduced burden upon schools and students and its impact on the likelihood of their continued willing participation.
■ each year group.	As is presently the case for the PSA targets, only certain year groups need their progress monitored.	
■ every year.	National monitoring surveys tend often to be undertaken only once every four years. This allows for different subject domains to be the focus of attention each	Since it is extremely hard to identify the extent to which improvement in an outcome measure genuinely and purely reflects improvement in the educational

Design principle	Design implication	Rationale
	year, for instance a four-subject programme might undertake one survey each year. However, where the system is intended to support the monitoring of progress towards national PSA targets, all subjects might have to be assessed in the same year and reported at the same time (before waiting another four years until the next survey series).	construct it is meant to reflect, there is very little understanding (internationally) of how much improvement in educational attainment ought actually to be expected from one year to the next. It is generally assumed, however, that large increases from one year to the next are unlikely (without radically changing the focus of education and, in particular, the amount of time devoted to teaching and learning specific elements). Surveys administered annually would be unlikely to reveal much in the way of change from one year to the next.

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

This version first published by The Office of the Qualifications and Examinations Regulator in 2008.

Office of the Qualifications and Examinations Regulator
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk